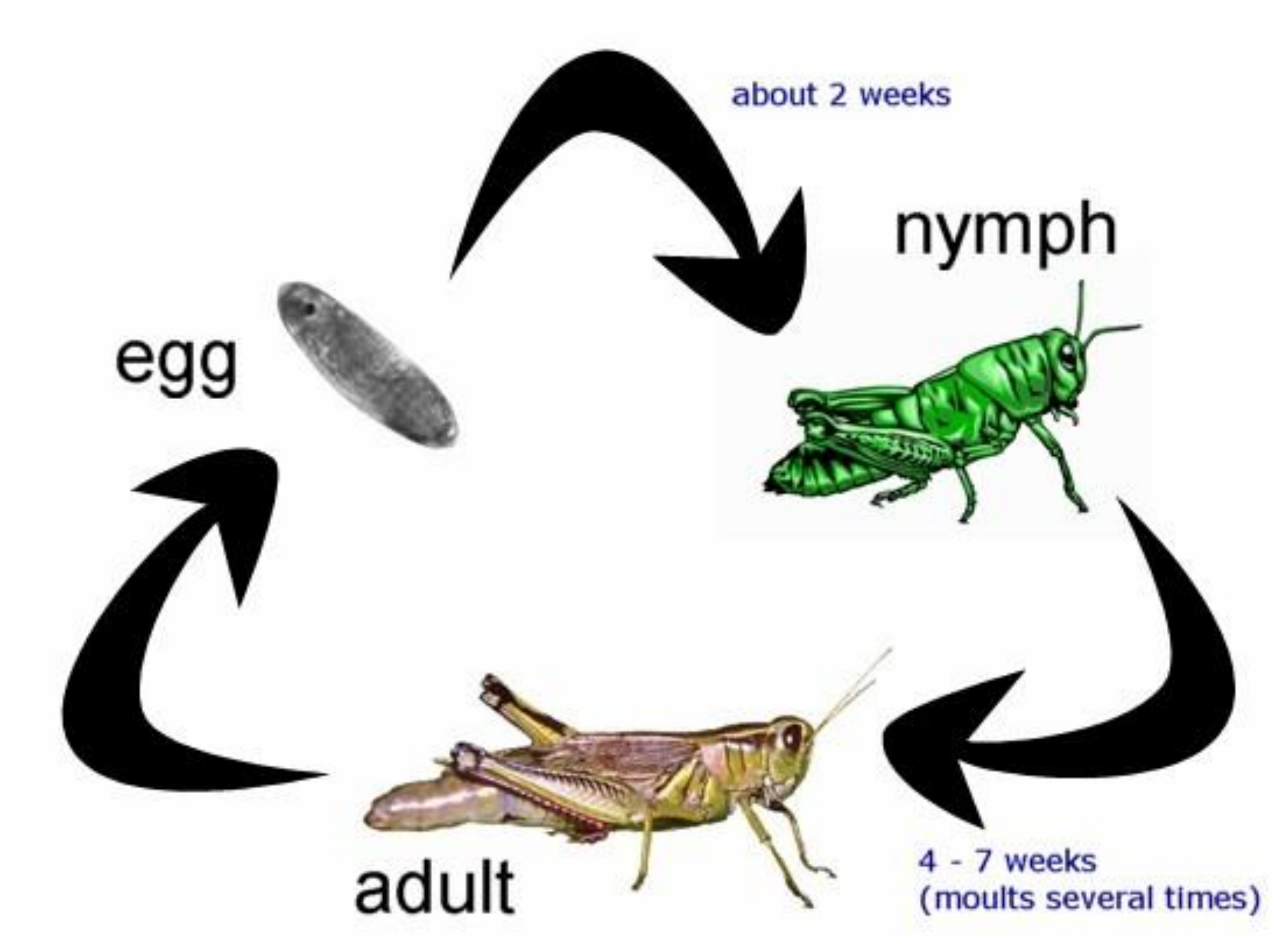
Using Stata to analyze size frequency in the life cycle of a Mexican desert spider

Irma Gisela Nieto Castañeda
María Luisa Jiménez Jiménez
Isaías H. Salgado Ugarte

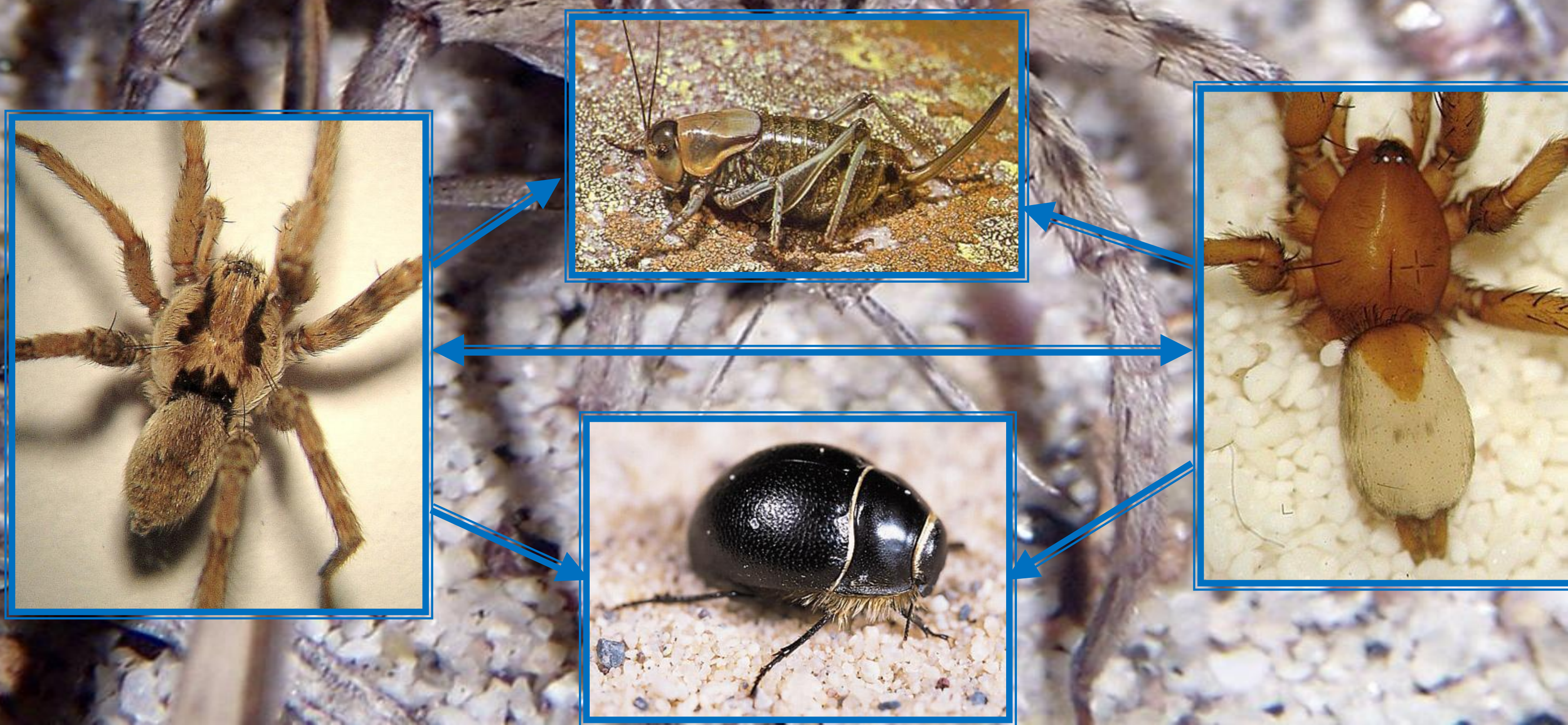
Life cicle in nature is particular and related with the living place and used resources for each organism



Spiders: abundant and diverse animals found in almost all environments (terrestrial and aquatic), short life cicles and very important in trophic webs

In deserts:

- Spiders are a very successful predator group
- They have morphological and physiological adaptations for avoiding extreme temperatures
- They forage any kind of animal that they can kill and eat



Syspira Simon, 1885

- These spiders live only in North American deserts
- They are nocturnal ground wandering spiders
- They represent almost 50% of all ground spiders of Baja California Sur, Mexico
- They are eaten by some rodents
- It is the first time that they are the subject of ecological studies

***Syspira tigrina* Simon, 1885**

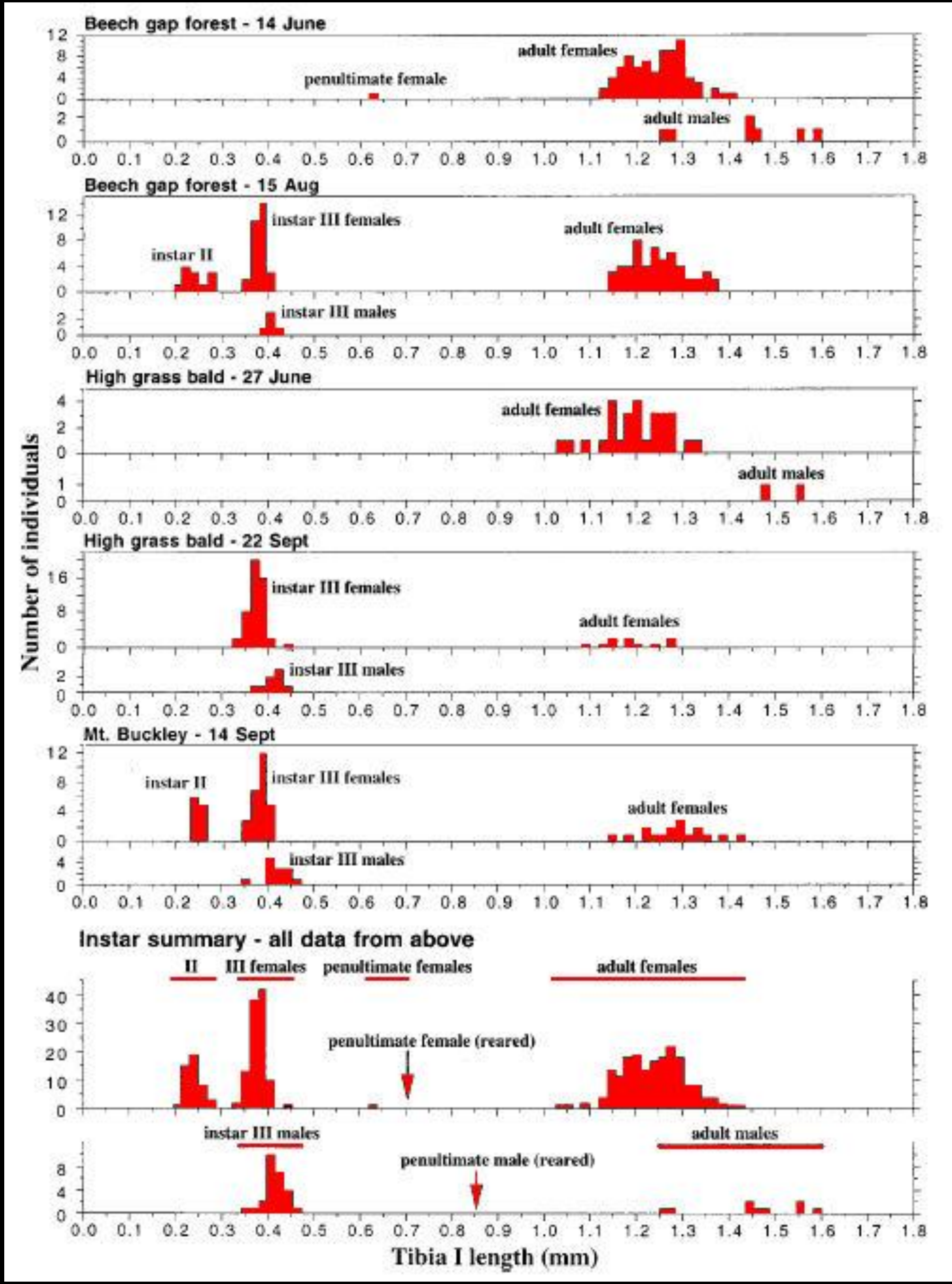
Photograph by IGNC

Life cycle

- It is unique for every species
- They have their own development and reproductive patterns
- Understand life cicles helps to clarify their biology and ecology

Life cycles in spiders

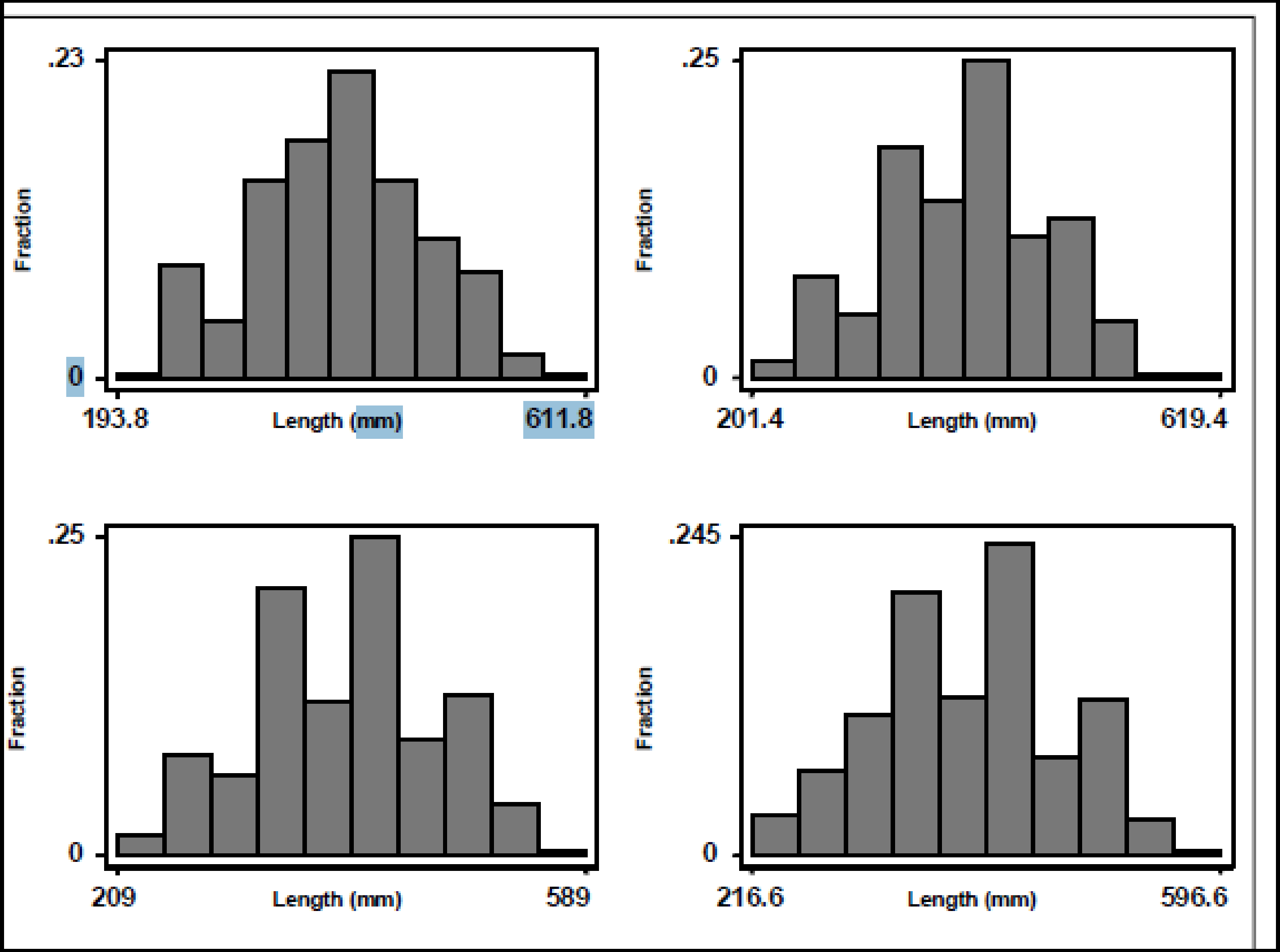
- These have been studied by two methods:
  - **Direct:** It keeps animals in captivity and follows their development and growth. It takes a lot of time and it is difficult to keep alive a representative sample of animals.
  - **Indirect:** Collects a big sample of animals during one or two years, measures every spider, and then finds a way to figure out how is the life cycle (found size classes or instars)



Example of the use of histograms to study the life cycle of spiders.

Each mode indicates a spider stage of development (instar)

Effect of origin on histograms: same data, same width, different origin; the histograms with shifted origins are bimodal, trimodal and tetramodal



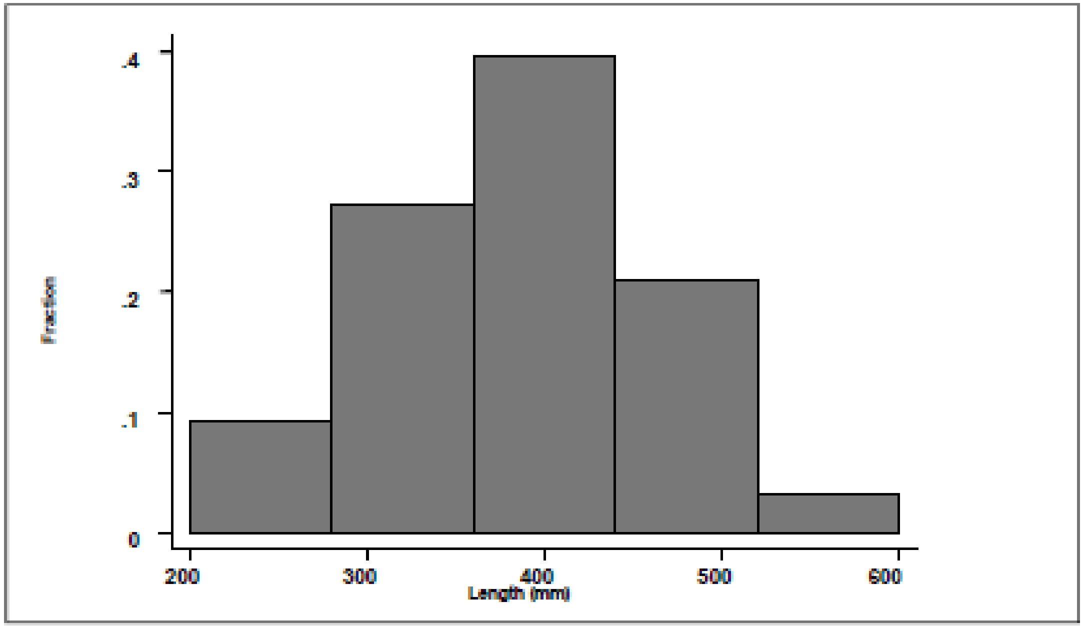
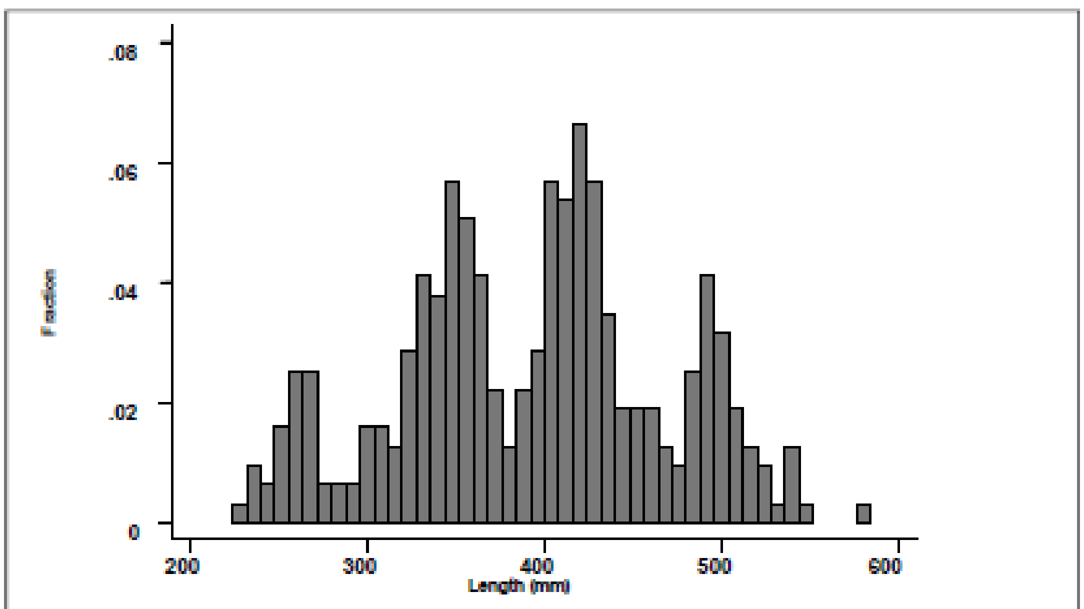


Fig. 2. Histogram with five bins for the coral trout length data.



Effect of number of intervals for the same data:

Few wide intervals: Simple (Gaussian- like distribution)

Many narrow intervals: Noisy multimodal distribution

Which one show the data distribution?

KDE's

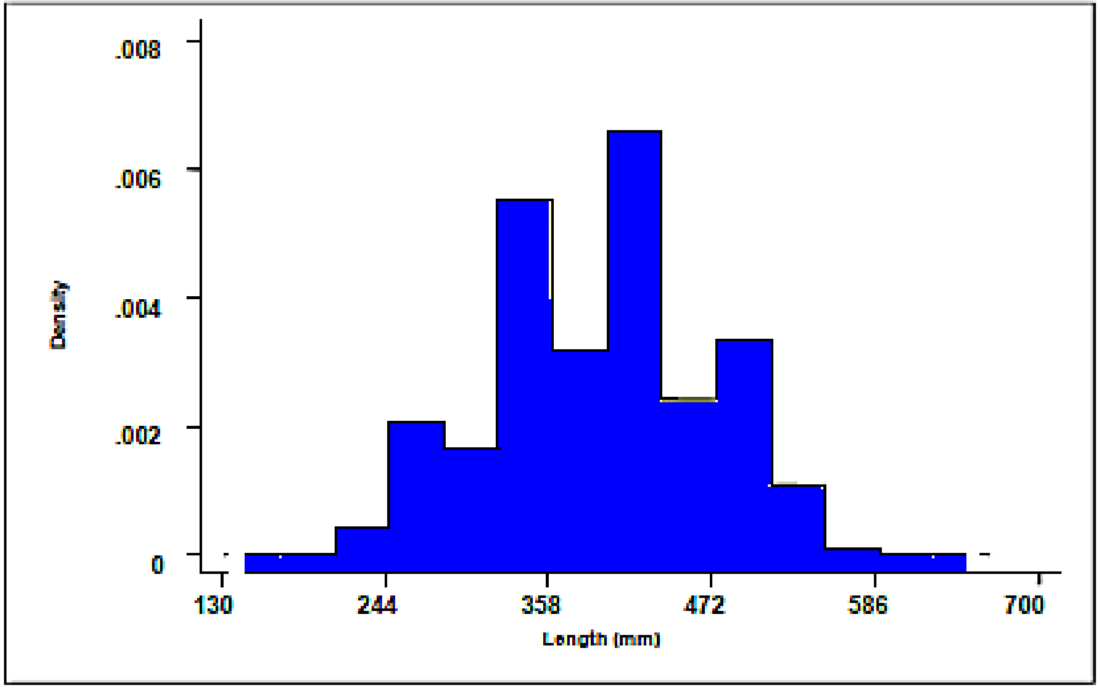
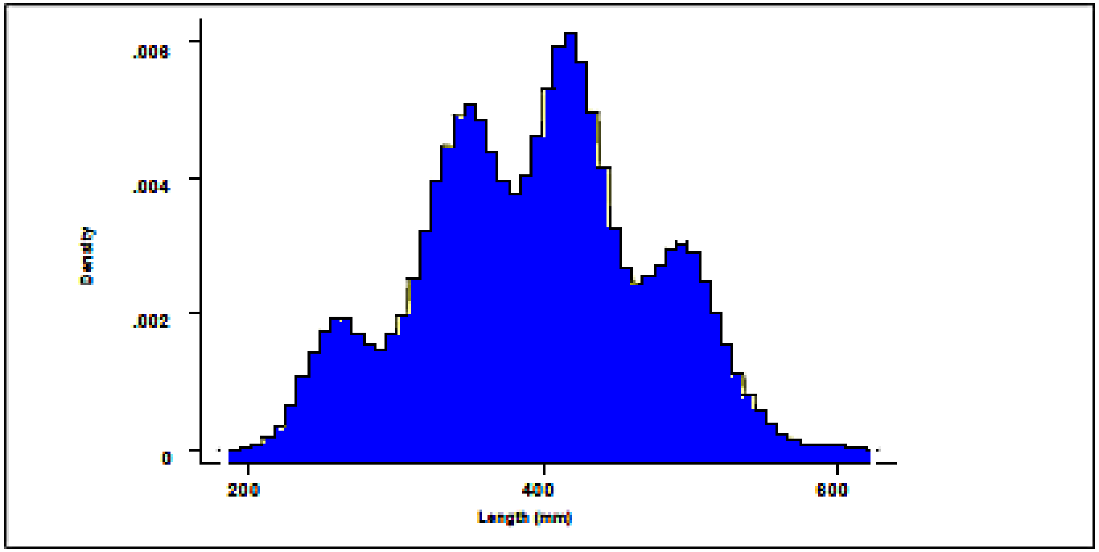


Fig. 4. Histogram with the optimal Gaussian binwidth (h = 38) and origin at 133 for the coral trout length data.



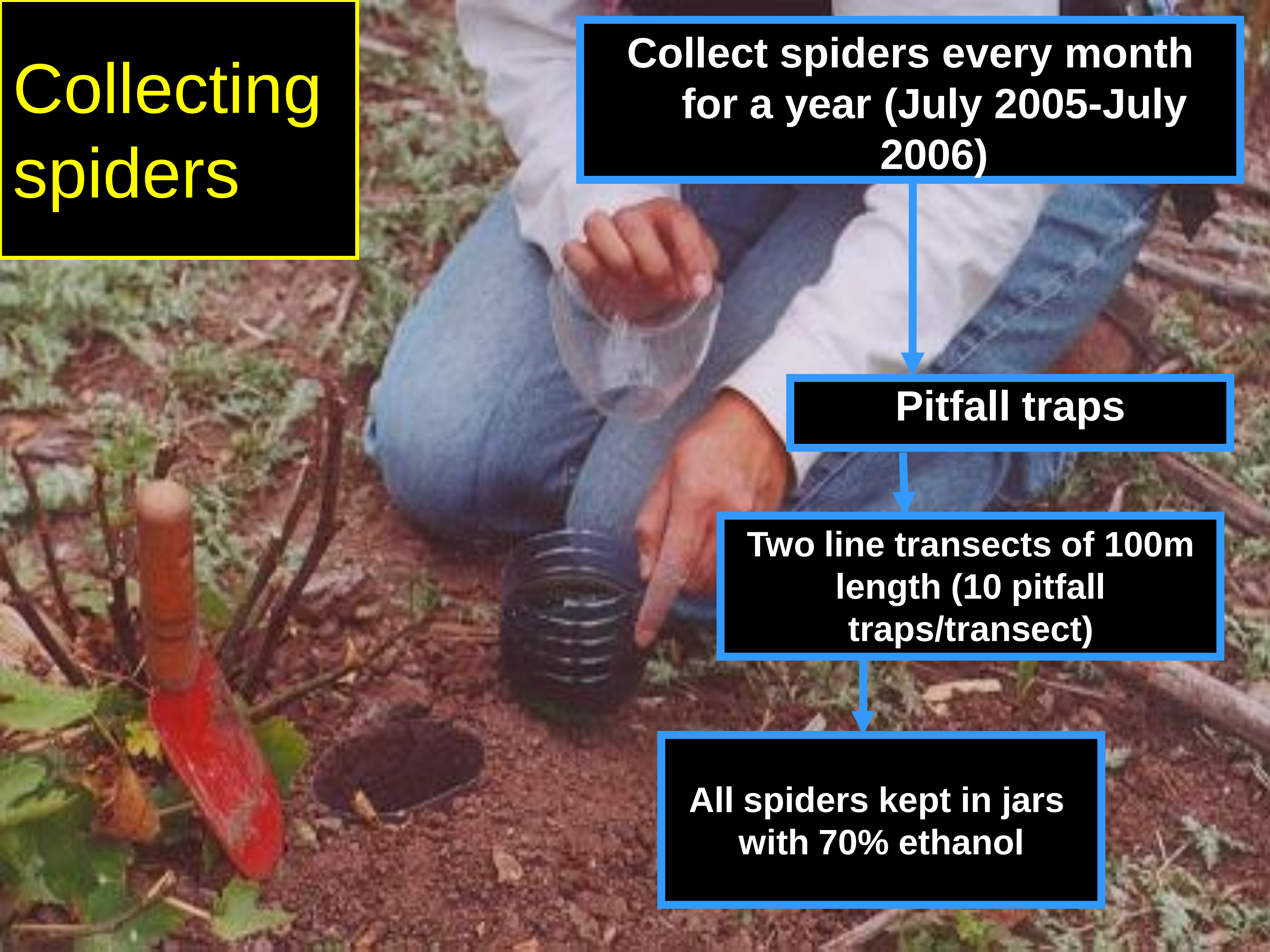
- Don't have the following problems:
  - Origin dependency
  - Discontinuity
  - Fixed interval width
- Helps to visualize:
  - Outliers
  - Skewness
  - Multimodality
- Every distribution has its own bandwidth

Objetive

- To describe the life cycle with the mixed size distribution of the *Syspira tigrina* species.

Hypothesis

- Because the EDK's method is efficient fot the analysis of data distribution, we must have a better approximation of how many size classes and their characterization are inside the life cycle of the species *S. tigrina*



Collecting spiders
Collect spiders every month for a year (July 2005-July 2006)
Pitfall traps
Two line transects of 100m length (10 pitfall traps/transect)
All spiders kept in jars with 70% ethanol

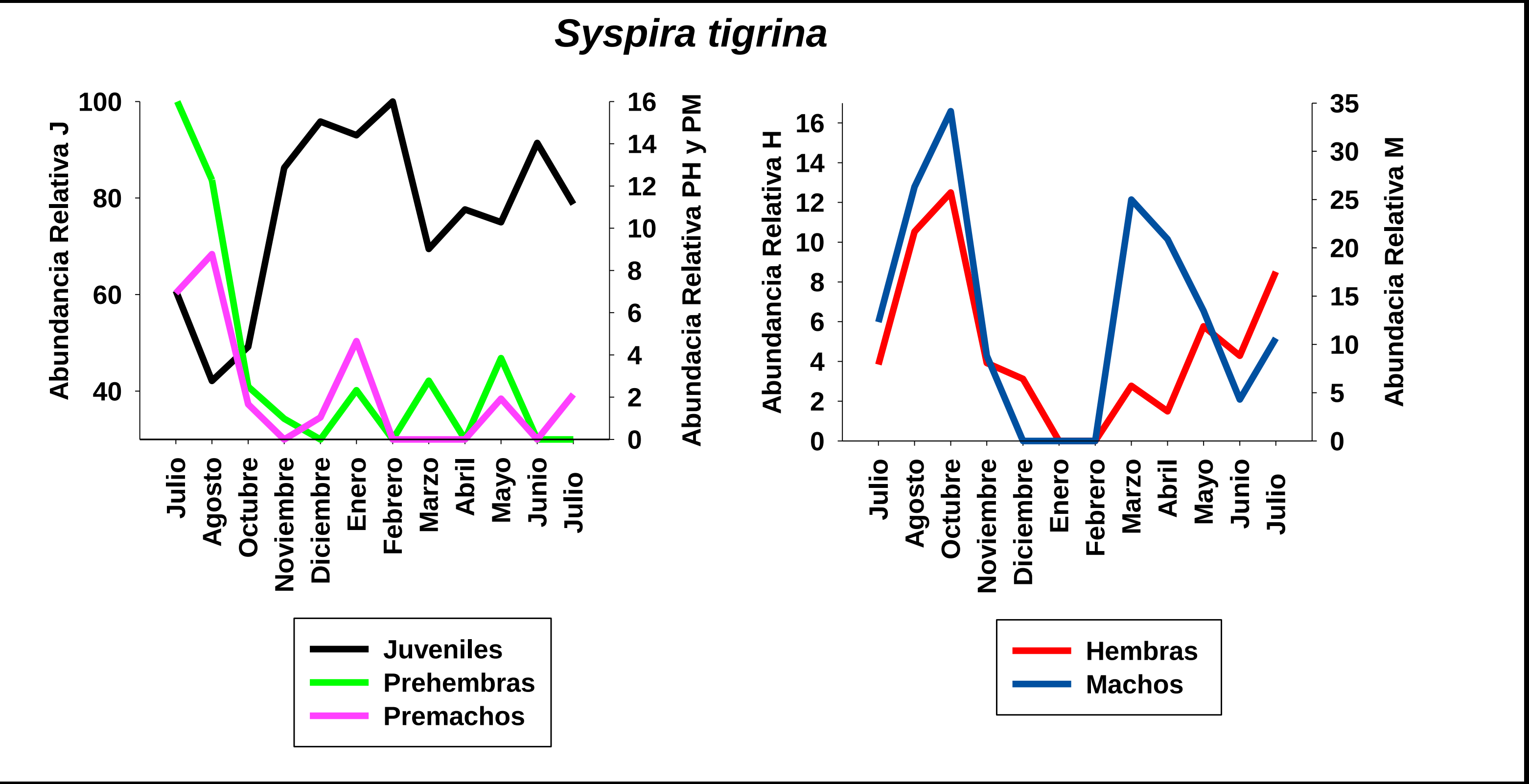
At Lab:

Clean every sample

Sort spiders

Identification of species

Measure the tibial length (TI)



When we found the highest number of adults (males and females) it corresponds with the lowest number of juveniles.

So we can figure out that the reproduction period should be before November, and then after this month the spiderlings start to emerge from cocoons

- We choose the bandwidth by the smoothed Bootstrap test of Silverman, and the Stata commands used were:
  - **bandw** (we took as reference the Silverman's "optimal" bandwidth and the Scott's oversmoothed bandwidth)
  - **critiband** (helps to find critical bandwidths)
  - **set seed** (to generate the pseudorandom numbers)
  - **boot bootsamb** (to generate the smoothed bootstrapped samples)
  - **silvtest** (smoothed bootstrap Silverman test)

An example of the command **bandw** use to analyze tibial length of *Syspira tigrina*. Oversmoothed and optimal bandwidths are indicated; were used as initial reference

```
. bandw t

Some practical number of bins and binwidth-bandwidth rules
for univariate density estimation using histograms,
frequency polygons (FP) and kernel density estimators
=====================================================================

Sturges' number of bins =                                    7.5236
Oversmoothed number of bins <=                               5.6877
---------------------------------------------------------------------
FP oversmoothed number of bins <=                            5.8347
=====================================================================

Scott's optimal Gaussian binwidth =                          0.1848
Freedman-Diaconis optimal robust binwidth =                  0.0726
Terrell-Scott's oversmoothed binwidth >=                     0.1104
Oversmoothed homoscedastic binwidth >=                       0.1969
Oversmoothed robust binwidth >=                              0.0944
---------------------------------------------------------------------
FP optimal Gaussian binwidth =                               0.2075
FP oversmoothed binwidth >=                                  0.2248
=====================================================================

Gaussian kernel (6)
=====================================================================
Silverman's optimal bandwidth =                              0.0442
Haerdle's 'better' optimal bandwidth =                       0.0521
Scott's oversmoothed bandwidth =                             0.1104
```

An example of the critiband command use: Critical bandwidths for one (0.1907) and two (0.1277) modes of tibial length are indicated

critiband t, bwh(0.192) bwl(.1260) st(.0001)m(40) nog

...

Estimation number = 12 Bandwidth = .1909 Number of modes = 1

Estimation number = 13 Bandwidth = .1908 Number of modes = 1

Estimation number = 14 Bandwidth = .1907 Number of modes = 1

…

Estimation number = 21 Bandwidth = .129 Number of modes = 3

Estimation number = 22 Bandwidth = .1289 Number of modes = 3

Estimation number = 25 Bandwidth = .128 Number of modes = 3

Estimation number = 26 Bandwidth = .1279 Number of modes = 2

Estimation number = 27 Bandwidth = .1278 Number of modes = 2

Estimation number = 28 Bandwidth = .1277 Number of modes = 2

Estimation number = 29 Bandwidth = .1276 Number of modes = 3

Estimation number = 26 Bandwidth = .1275 Number of modes = 3

An example of the silvtest command use:

The recommended bandwidth is obtained by calculating the midpoint of all the bandwidths with three modes (from 0.2998 to 0.1112) = 0.2055

```
. silvtest ysm _rep, cr(0.0757) m(40) nuri(1) nurf(500) cnm(3)

Critical number of modes =      3

--------------------------------------------------------------------

Pvalue =              291 / 500 =   0.5820

. di (0.2998+0.1112)/2
.2055

. clear
```

Two examples of Tables with the Silverman test results. A *P* value equals or larger that 0.4 indicates the number of modes with statistical significance

November 2005

N = 88

Modes	Critical Bandwidth	Seed	Bandwidth	*P*
1	0.2997	832467	0.23665	0.0000
2	0.1112	737895	0.20635	0.4200

December 2005

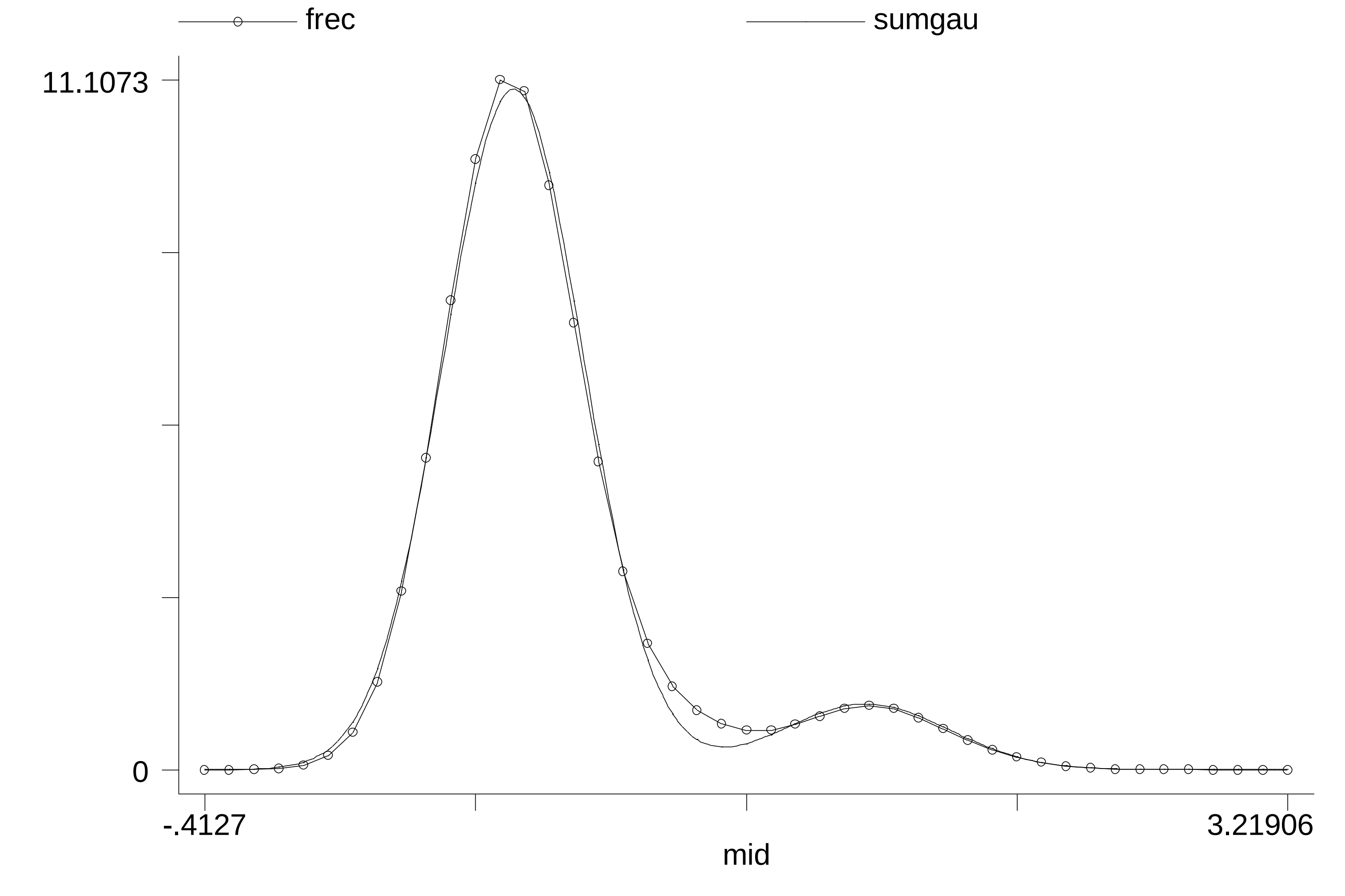
N = 92

Modes	Critical Bandwidth	Seed	Bandwidth	*P*
1	0.2168	82455	0.1636	0.0400
2	0.1084	75757	0.16265	0.2400
3	0.0637	13571	0.0861	0.3400
4	0.0538	95827	0.0588	0.2200

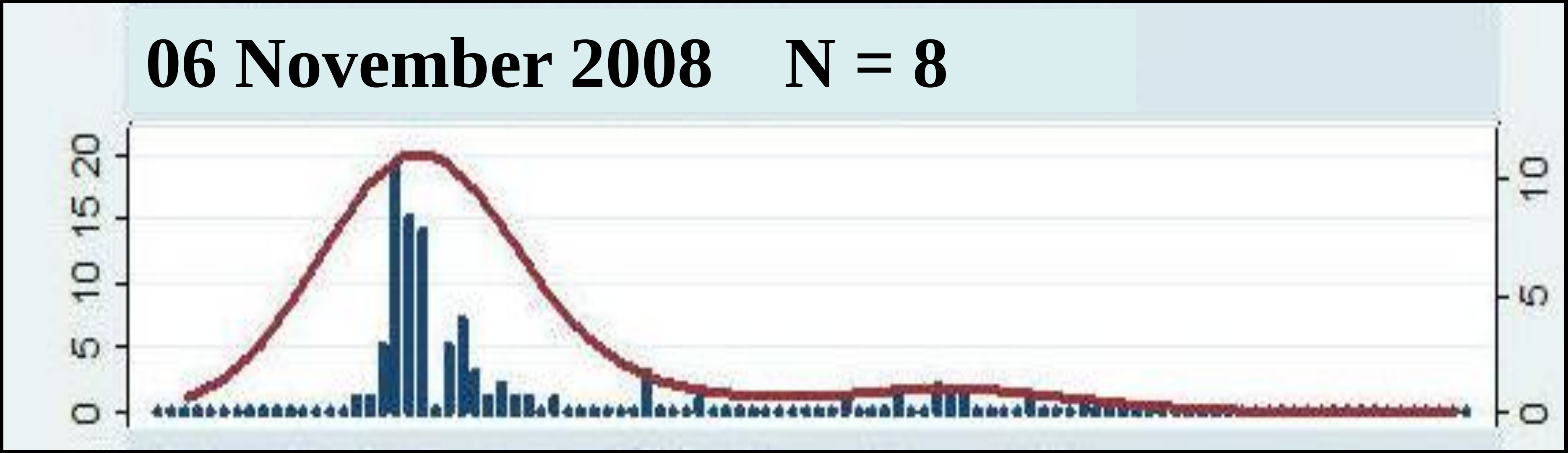
Gaussian components with their parameters obtained by the Bhattacharya's method representing the stages from twelve samples of *Syspira tigrina*

Juveniles					
	Gaussian Component				
Date	Number	Midpoints range	Mean	Standard deviation	Size
26 July 2005	1	11-23	0.7710	0.2530	22
	2	21-27	1.6270	0.4041	58
	3	40-43	2.2967	0.1752	1
27 August 2005	1	8-14	0.5874	0.2862	31
	2	20-27	1.5863	0.4425	20
4 October 2005	1	5-14	0.6051	0.1618	36
	2	17-21	0.9644	0.1787	15
	3	31-35	1.5241	0.3452	10
	4	38-42	1.9755	0.1846	2
6 November 2005	1	10-17	0.6228	0.2343	78
	2	27-33	1.8053	0.2804	9

KDE with the sum of two Gaussian components (bimodal distribution)

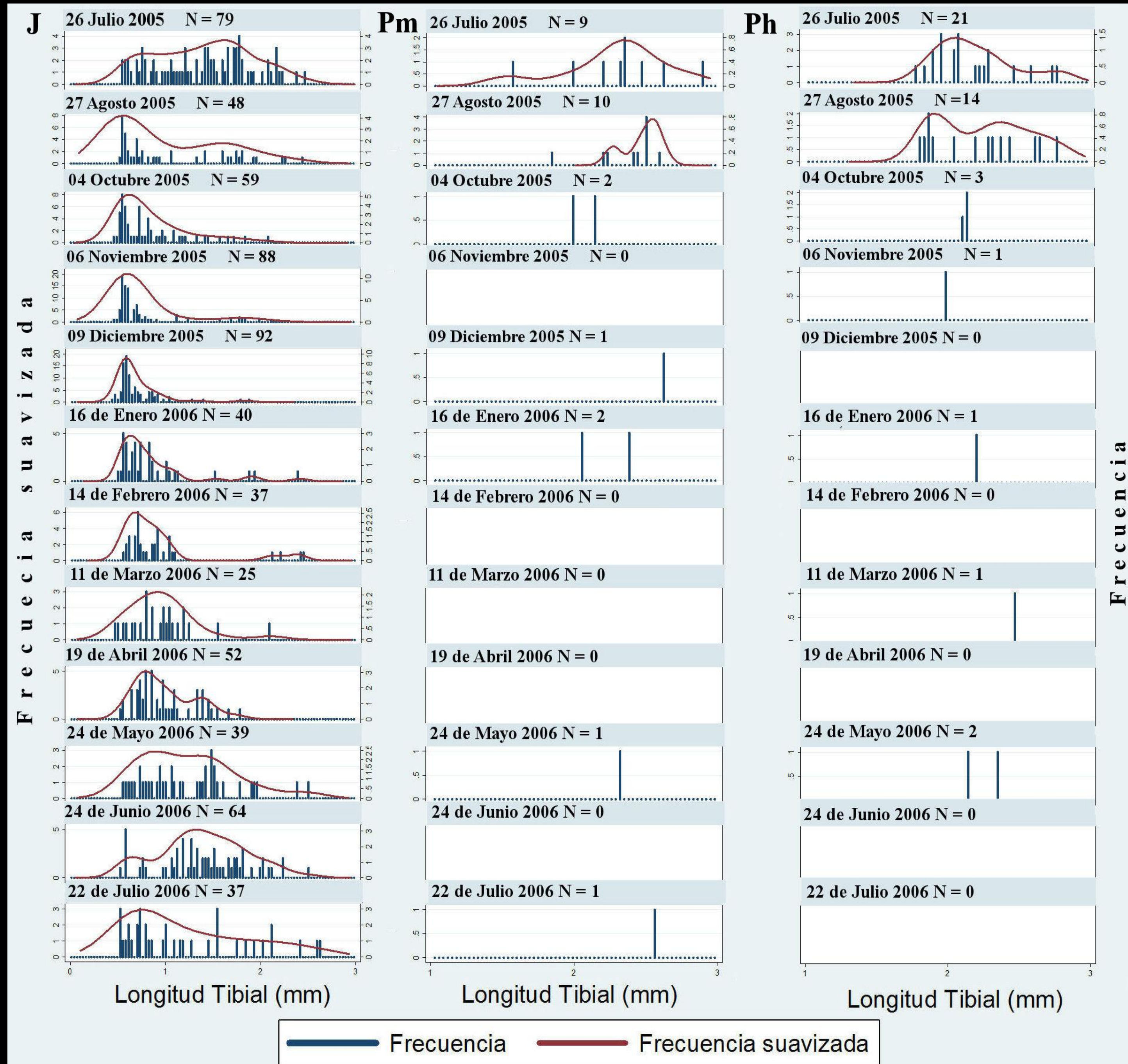


KDE+ histogram
November 6th, 2005



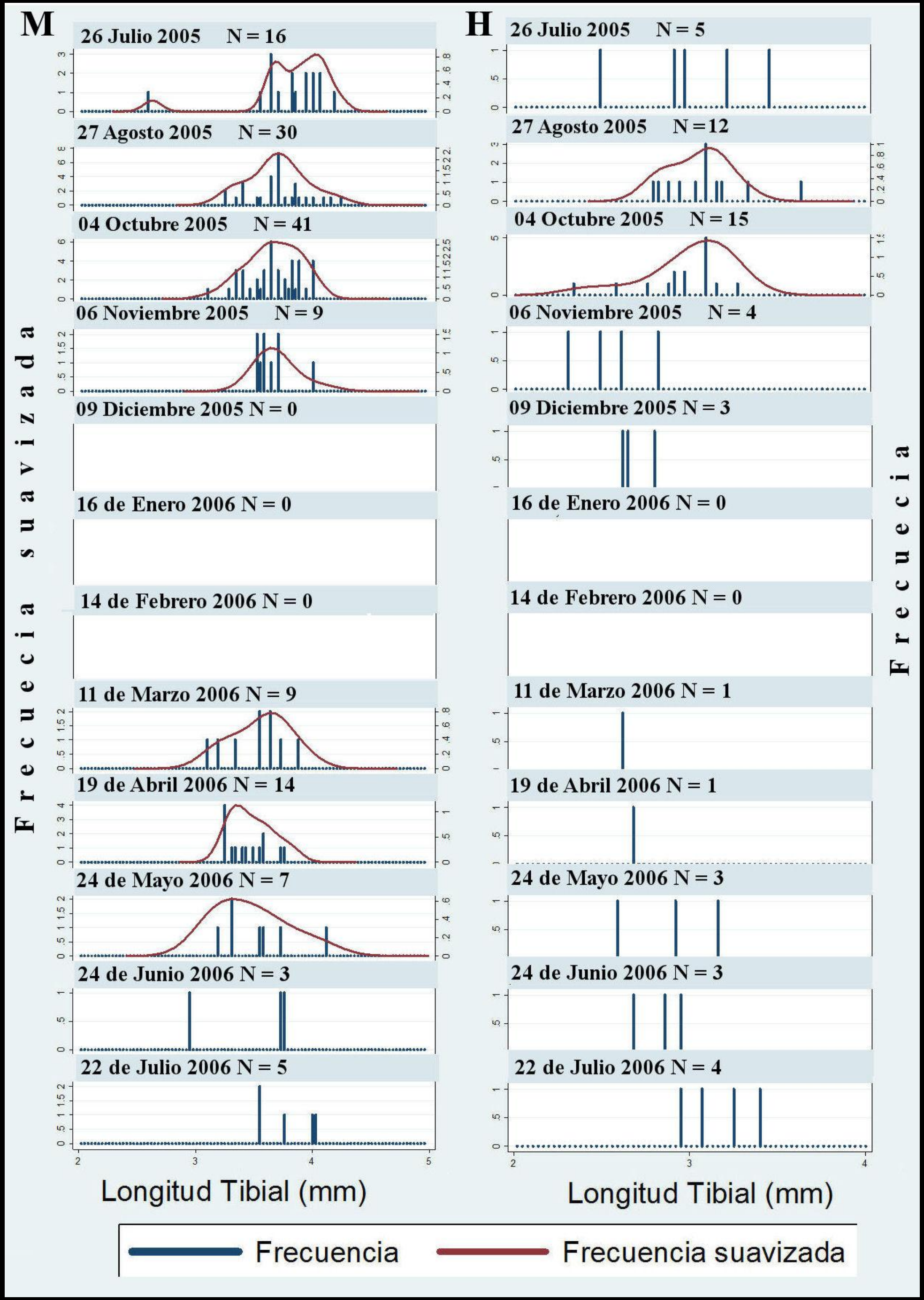
Total results:

KDE’s and histograms for the juveniles (J) penultimate-males (Pm) and penultimate females (Ph) applied to analyze the tibial length



Total results:

KDE's and histograms for the Males (M) and Females (H) applied to analyze the tibial length



Conclusions

- We recommend to analyze size classes instead instars, because sometimes there are no relationship between age and size.

- Identified size classes should mean that all organism from the same group should use resources in a similar way

- The EDK's are a very good option (and better than histograms) to find and characterize size classes of mixed distributions such as those from *S. tigrina* samples

Acknowledgments

- CONACyT (Mexico)
- Miguel Correa and Carlos Palacios for field work support
- PhD tutorial committee
  - Dra. Maria Luisa Jimenez
  - Dra. Carmen Blázquez
  - Dr. Guillermo Ibarra
  - Dr. Yann Henaut
  - Dr. Frederick A. Coyle

Remember spiders are so famous in deserts that even camels want to look like them